# K740) Assignment Three: Association Rules

### Submission due Thursday, March 13, 2023 at noon.

Please submit this report (50% of the grade) bundled with the Jupyter notebook (50% of the grade) you used to generate results.

## The problem:

You have received 1000 transactions from a small local grocery store and are going to provide some suggestions to increase sales and improve customers’ shopping experience.

For completing assignment three, please perform the following tasks and answer the questions:

## Downloading materials

1. Download and save the transaction file: (groceryStoreTransactions.csv)
   * The file contains 1000 grocery shopping transactions.
2. Use codes from Jupyter notebooks (Association Rules and Linear Regression sessions) to complete this assignment. Please:
   * Change variable names, directory paths, and values for evaluation metrics using the information in each question.
   * Remember that column names in groceryStoreTransactions.csv file are different from column names in the fake transactions we created in the class.

## Summarizing the transactions **(15%)**:

1. How dense is the dataset you have received? **(2%)**

**The density of dataset is just 34.66%, which means that 65.34% are zero(non value). Therefore, the dataset on hand is sparse.**

1. How many items does the store sell? **(3%)**
   * Please provide the list of items.

**The store sells 20 items. They are: 'Bordeaux', 'Coke', 'Heineken', 'apples', 'artichoke', 'avocado', 'baguette', 'bourbon', 'chicken', 'corned beef', 'crackers', 'ham', 'herring', 'ice cream', 'olives', 'peppers', 'sardines', 'soda', 'steak', 'turkey'.**

1. What is the average number of purchases per customer? **(3%)**
   * Hint: Use dataset’s density and number of items to calculate this.

**20\*0.3466 = 6.93 round up to 7 items per purchase.**

1. What are the 5 most purchased items? What fraction of customers have purchased these items? **(5%)**

**The 5 most purchased items are Heineken (600 times), crackers (488 times), herring (485 times), olives (472 times), and bourbon (402 times).**

**The fraction of customers who purchased Heineken is 30% (600/2000)  
The fraction of customers who purchased crackers is 24.4% (488/2000)**

**The fraction of customers who purchased herring is 24.25% (485/2000)**

**The fraction of customers who purchased olives is 23.6% (472/2000)**

**The fraction of customers who purchased bourbon is 20.1% (402/2000)**

1. How many possible rules can you create with the items the store has? **(2%)**

**2^20 – 1(cannot walk away without purchasing any item) = 1048575 possible rules.**

## Generating rules **(30%)**

1. Generated rules should be supported by at least ten percent of transactions and provide more than 50% confidence. **(10%)**
2. How many rules have been generated (in general or with two, three, four, and five items)? **(10%)**

**581 rules in general; 52 rules with two items; 215 rules with three items; 231 rules with four items; and 83 rules with five items.**

1. Use summary statistics to summarize the generated rules (with different lengths) regarding their support, confidence, and lift. **(10%)**

**The mean of support for each lens of rules tells that the rule with more items has less support on average, which makes sense, as the pattern with 2 items (buying 2 items) should be more frequent than the pattern with 3 items and so on.**

**Whereas the confidence will increase if the rule covers more items. It is also expected, as longer rules will increase the accuracy.**

**The lift will also increase if the rule covers more items. The lift measures how much more often the antecedent and consequent occur together than expected under independence. In this case, life means the chance A will be purchased when B is purchased at the same time. When the purchase covers more items, the increase in confidence may offset the decrease in support, making the rule more interesting and leading to an increase in lift.**

## Plotting the rules **(20%)**

1. Use a scatter plot matrix (*from Linear Regression Jupyter notebook*). Use support, confidence, and lift for the dimensions, and rule length for color. **(10%)**
   * Hint, you need to add rule length to your rule data frame.
2. What pattern do you observe among the length of rules, support, confidence, and lift in the scatter plot matrix? **(10%)**

**The first pattern: the longer the rule, the less support and more lift it will have. And the majority rules tend to have support less than 0.15.**

**The second pattern: it is hard to tell the effect of the length of rule on the confidence although they should have a positive relationship, it is hard to see the difference in the scatter plot.**

**The third pattern: for the same length of rule, when confidence increase, the lift will also increase.**

## Making Recommendation **(35%)**

1. One of your suggestions is to bundle some items and offer competitive prices (or place items closer to each other). What are your choices of two sets of items for bundling or placing, and why? **(20%)**  
   Hint: Select a subset of rules with a high confidence ratio (start with 95%) and high lift (start with 3.5)

**If we want to bundle sales, we have to choose to bundle the products that have a higher likelihood of being purchased together. After raising the confidence ratio to 95%, we can say all the rules we get will be highly likely to happen. So based on the rank of lift (the probity of the antecedent and consequent happening together),** I **suggest bundling Heineken, sardines, Coke with ice cream, chicken and ham, herring, turkey together with corned beef, olives. Both bundles have a confidence of more than 96% and lift higher than 4.1.**

1. If you have 500 customers per day, how many bundles do you estimate you will need on a daily basis? Why? **(15%)**

**We will need 58 (500\*0.116) bundles of Heineken, sardines, Coke together with ice cream, chicken, and 50.5 (500\*** **0.101) bundles of ham, herring, turkey together with corned beef, olives.** **Support is a measure of how often a particular combination of items appears together in a transaction or dataset. If all 500 customers purchase some product, there will be 500 transactions each day. In these transactions, the number of transactions of each bundle then will be 500 times the means of the support calculated in the model. Therefore, we should need that amount of inventory.**